SC24 HPC-AI Market Update - Cloud

SC24 Breakfast Briefing
November 2024

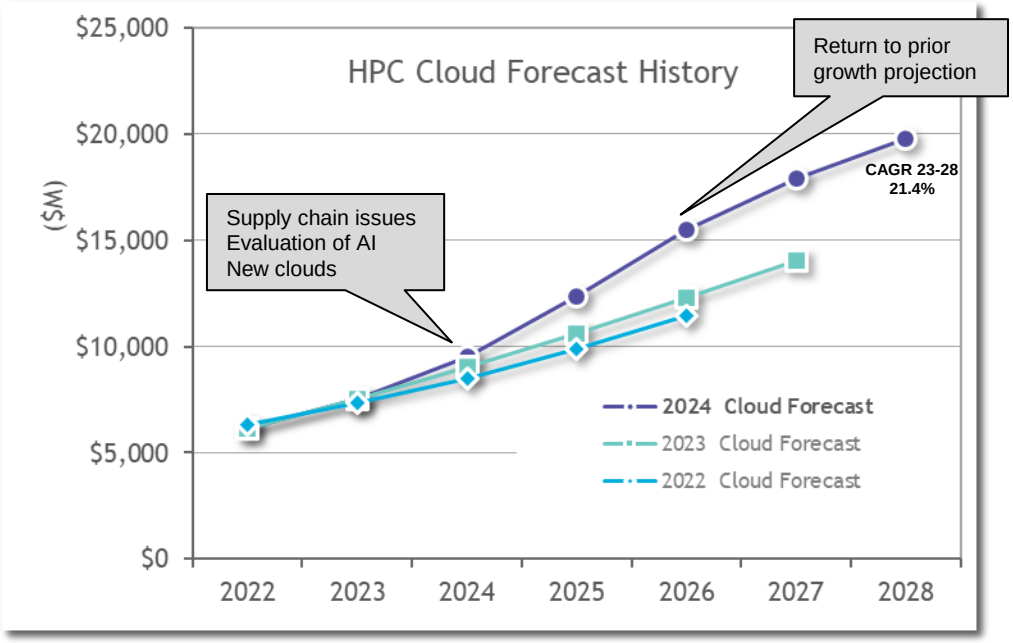
www.HyperionResearch.com
www.hpcuserforum.com

Mark Nossokoff

HPC/AI Cloud Forecast

Cloud revenue expected to approach $20B by 2028

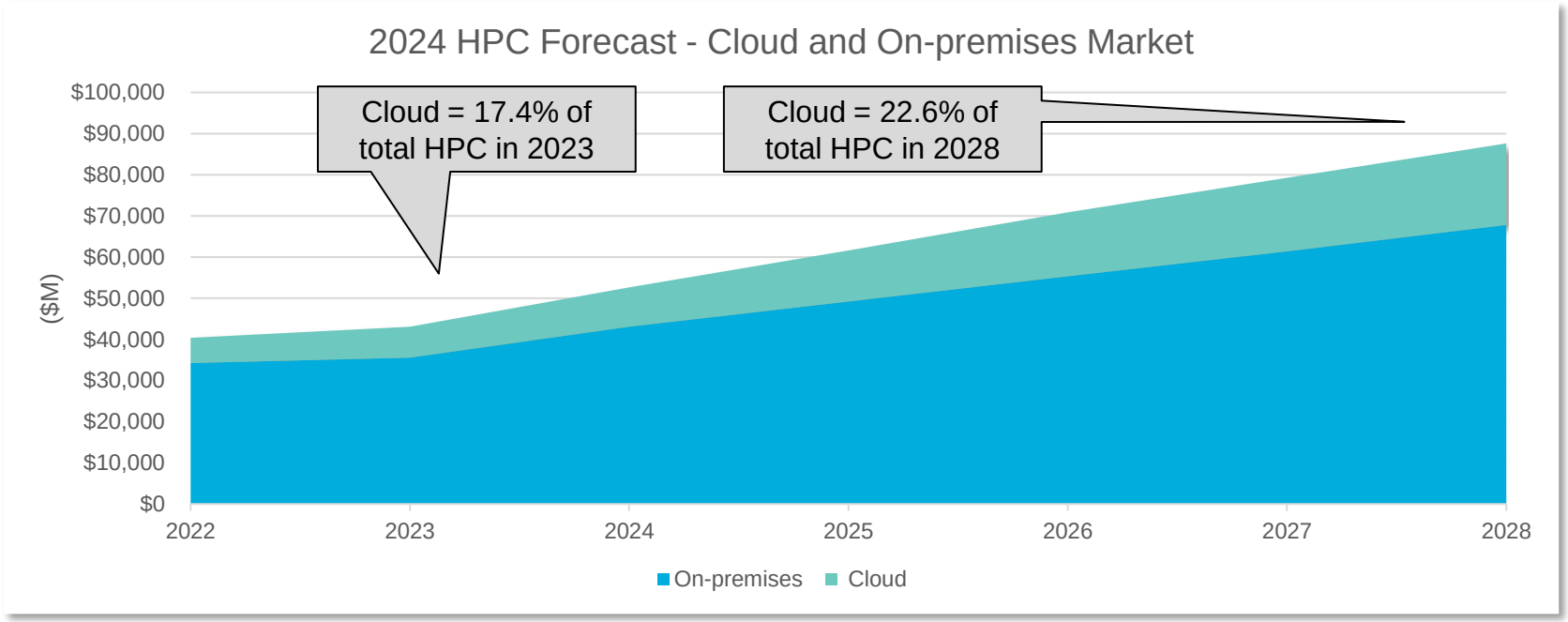
- **Global HPC/AI buyers around the world continue shifting portions of their on-premises budgets to spending in the cloud**
- **Increasing growth rate for 2024, returning to prior projected growth in 2026**
- **Primary growth drivers:**
  - Supply chain issues
  - Access to latest GPU technology
  - Experimentation and evaluation of AI workflow and infrastructure needs
  - Emergence of sovereign clouds



- **Forecast has steadily increased over previous forecasts**

The Total HPC/AI Market: On-Prem and Cloud Computing

Total HPC-AI exceeds $87B in 2028



	2022	2023	2024	2025	2026	2027	2028	23-28 CAGR
Cloud	$6,132	$7,516	$9,540	$12,376	$15,519	$17,892	$19,804	21.4%
On-Premises	$34,250	$35,573	$43,054	$49,223	$55,315	$61,390	$67,805	13.8%
Total	$40,382	$43,089	$52,594	$61,599	$70,834	$79,282	$87,609	15.2%

Cloud Providers' Responses to Demand

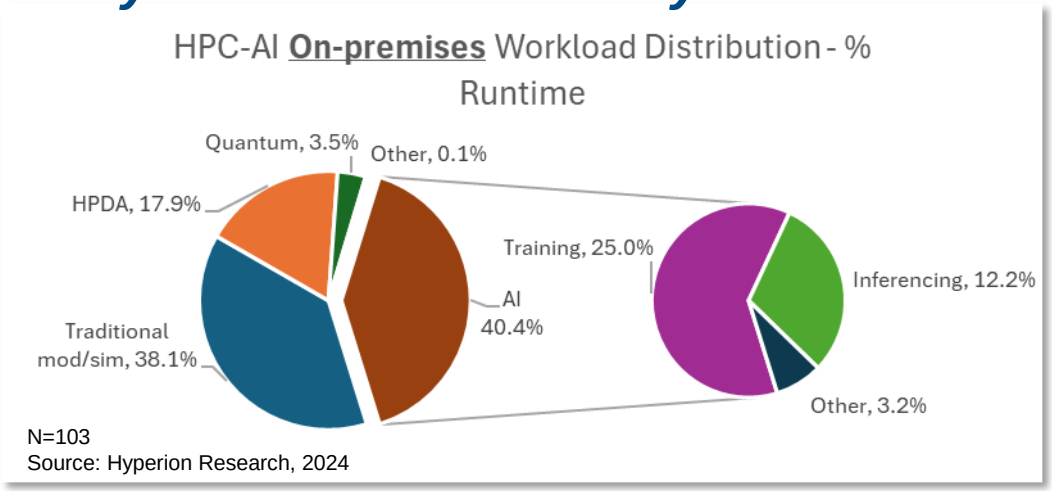
Rising cloud utilization driven almost entirely by AI adoption

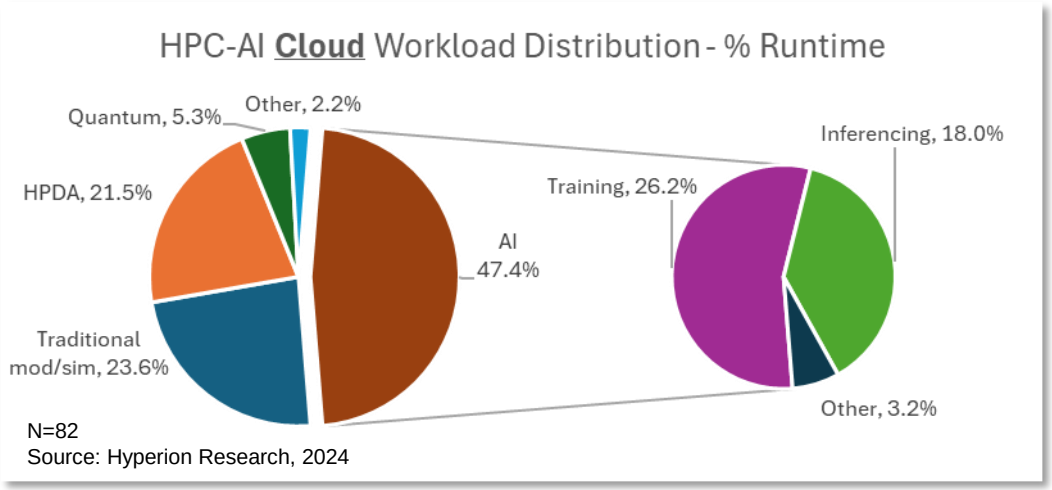
- **Extensive investments by cloud providers**
  - Application-specific processors and accelerators
  - New data center buildouts
  - Expanded HPC focus from CSPs
    - AWS released Parallel Compute Service
    - Google GA's Parallelstore
- **AI and GPU clouds**
- **Sovereign clouds**
  - Driven by local legal requirements, cost constraints, and tax & privacy policies, a shift to smaller, more adaptable data centers could result
  - CSP's ability to dynamically shift load demand between geographic regions and national borders may be impacted
    - Local provisioning to become more critical
  - What is the correct metric to key in on to retain "sovereignty"?
    - Where data is created and where it can be moved, or…
    - …security and governance of users who can access it, wherever it may be stored?

HPC-AI Workload Distribution by Environment - % Runtime

Of all your workloads in your HPC/AI/HPDA on-premises/cloud environments, please distribute your utilization time by the following:

- **AI identified as the primary workload based on runtime**
- **AI approaching 50% of the workload runtime in the cloud**
- **Traditional mod/sim runtime is 61% greater on-prem than in the cloud**

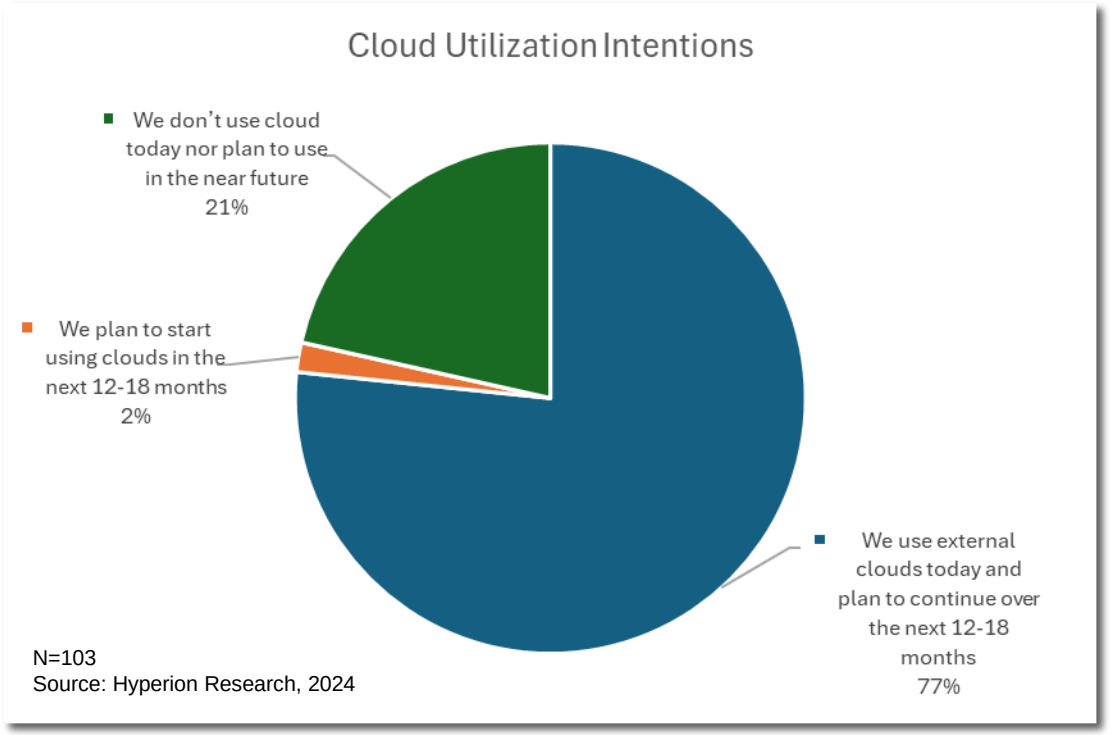




Cloud Utilization for HPC-AI Workloads - Intentions

Are you using or planning to use external cloud resources for any of your HPC, AI, big data, or quantum workloads?

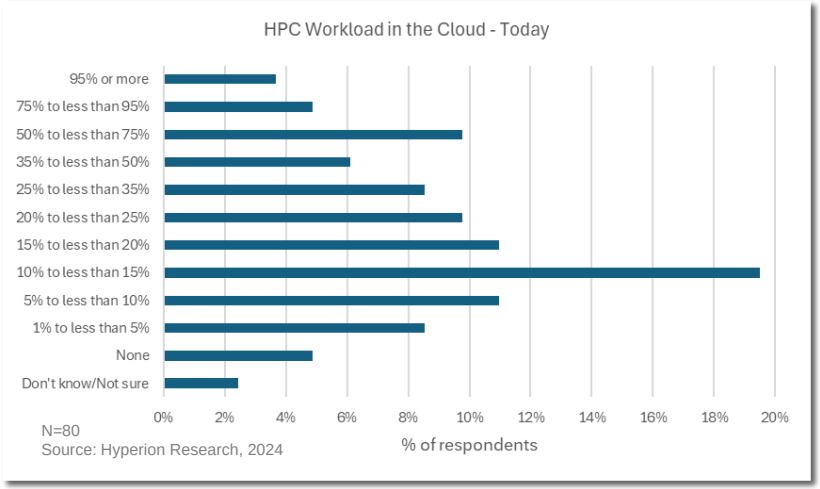
- **70% of respondents currently use or intend to use cloud within the next 12-18 months**
- **21% don't use the cloud or intend to use the cloud within the next 12-18 months**

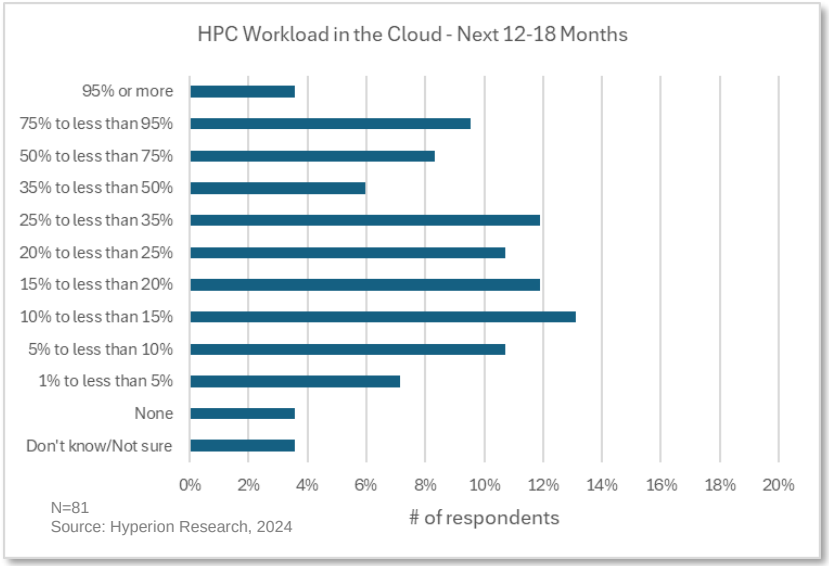


Cloud Utilization for HPC-AI Workloads - % Runtime

Based on overall runtime, approximately what percentage of all your HPC-AI workloads are run on external clouds TODAY/12-18 months?

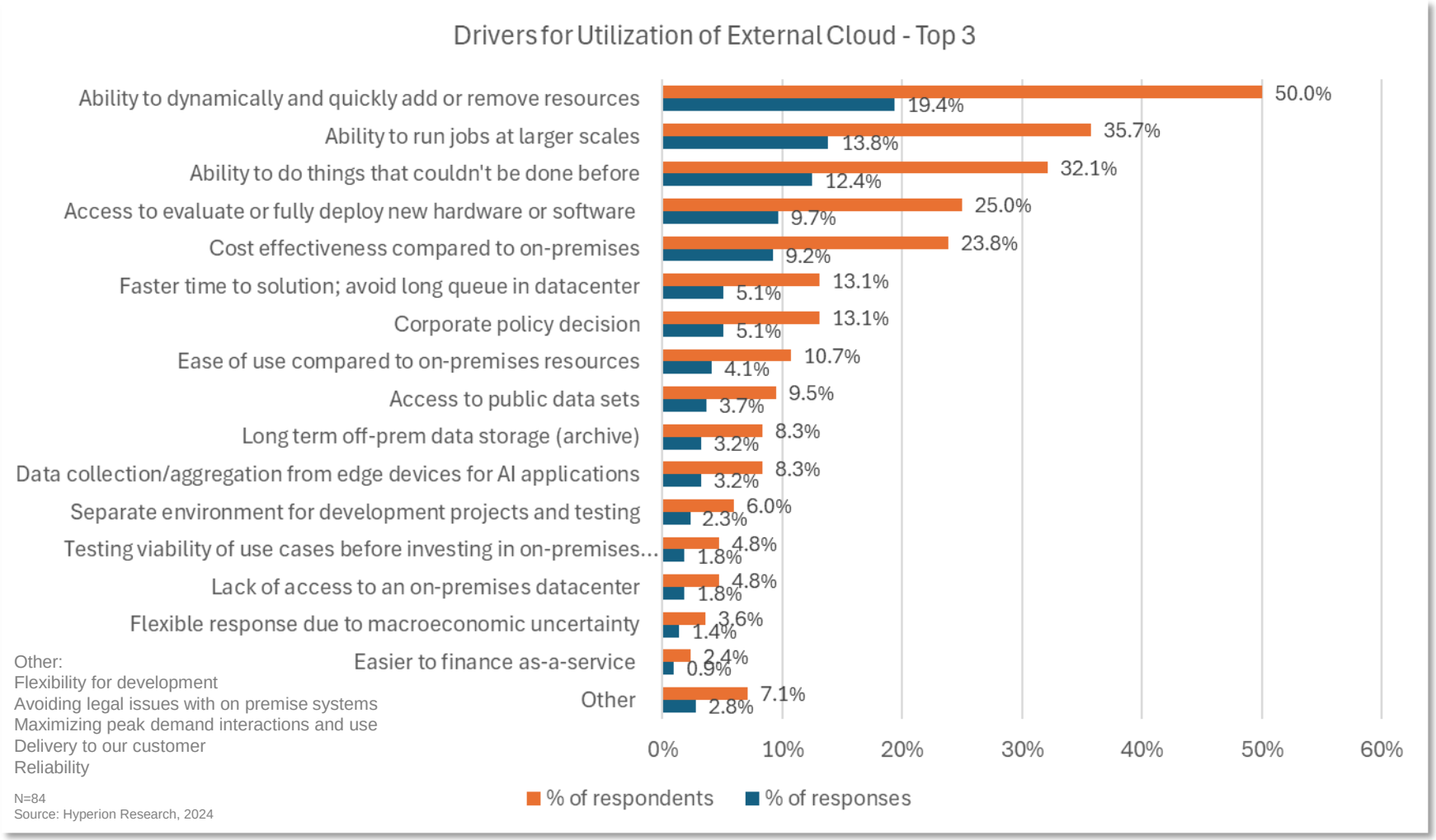
- **Respondents indicate an increase in application runtimes performed in the cloud (weighted average)**
  - Today: 27.2%
  - 12-18 months: 31.1%
- **Respondents running 50% or more of their application runtimes in the cloud growing**
  - Today: 18.8%
  - 12-18 months: 22.2%





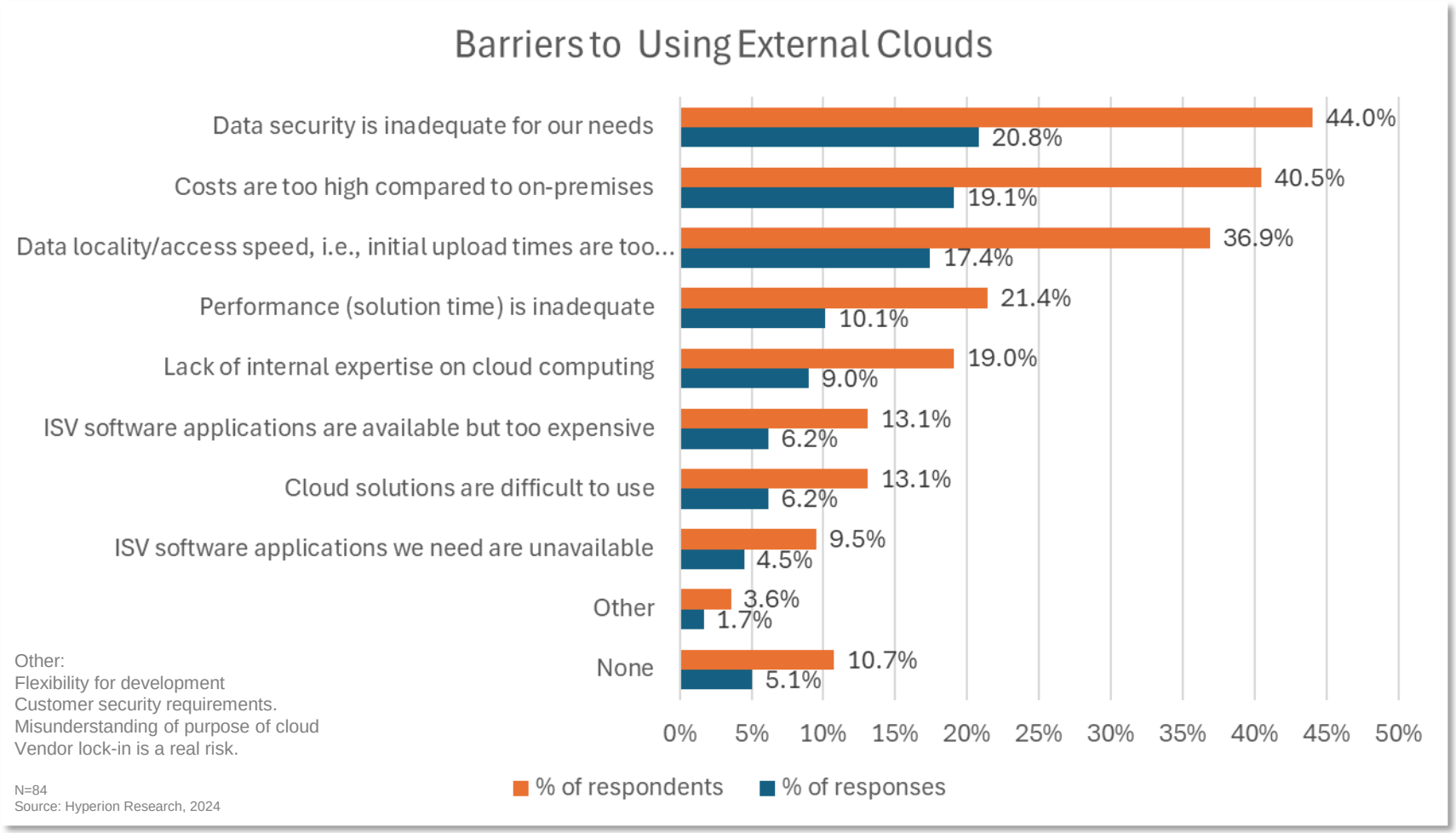
Drivers for External Cloud Adoption

Which of the following is a reason you use/plan to use external clouds for HPC? Please select up to 3.



Barriers for External Cloud Adoption

Which of the following do you consider to be barriers to increasing EXTERNAL cloud use for HPC workloads? Please select up to 3.



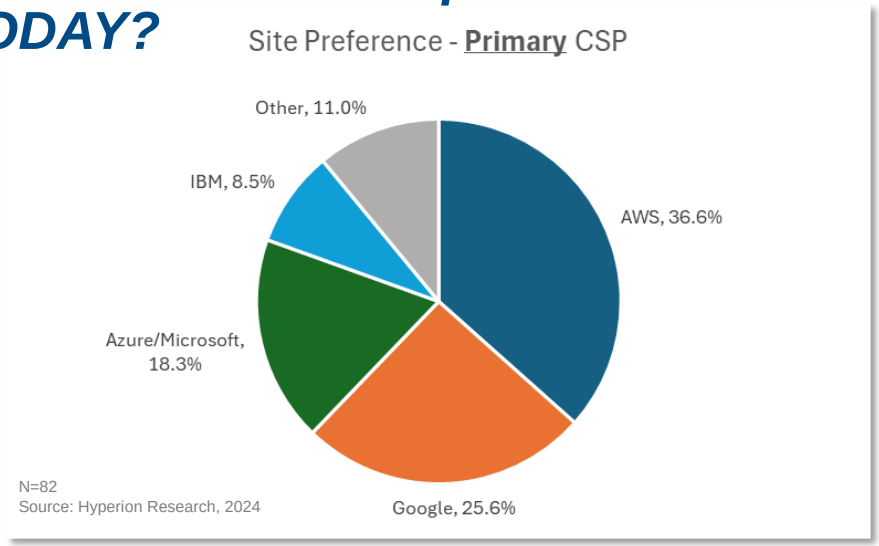
Other:
Flexibility for development
Customer security requirements.
Misunderstanding of purpose of cloud
Vendor lock-in is a real risk.

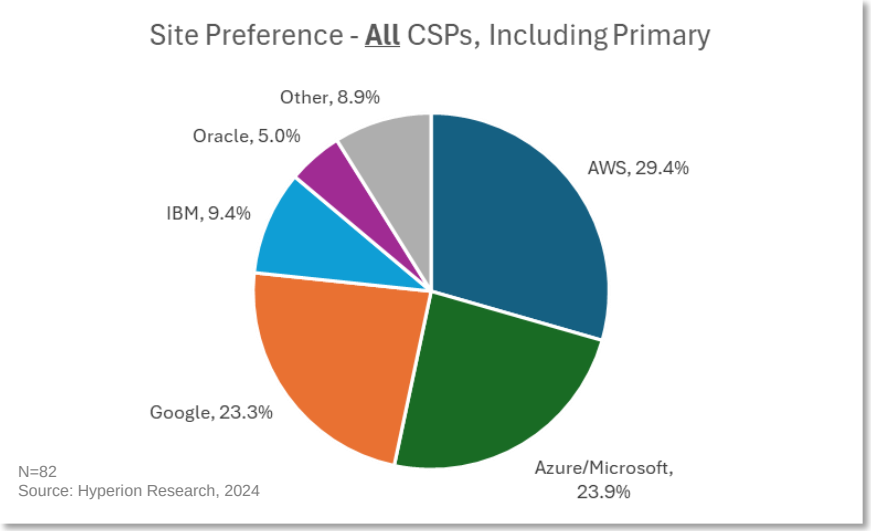
N=84
Source: Hyperion Research, 2024

CSP Preferences – Primary vs. All

Who is your PRIMARY cloud provider / ALL cloud providers for your HPC/AI/HPDA workloads TODAY?

- **AWS the preferred primary CSP among respondents**
- **Google the 2nd most preferred primary CSP**
- **Microsoft the 3rd most preferred primary CSP, but rises to 2nd when considering all CSPs**
  - 180 total responses for CSPs utilized
  - ~2 CSPs per site

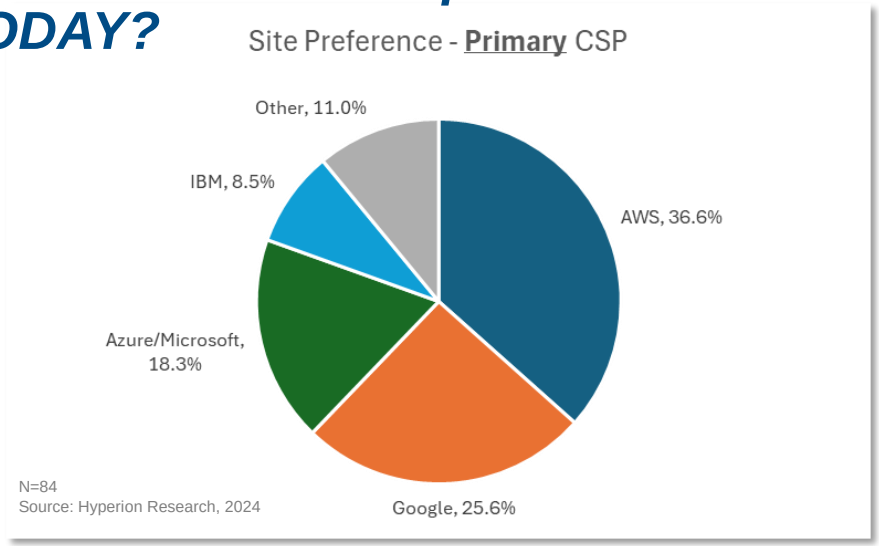


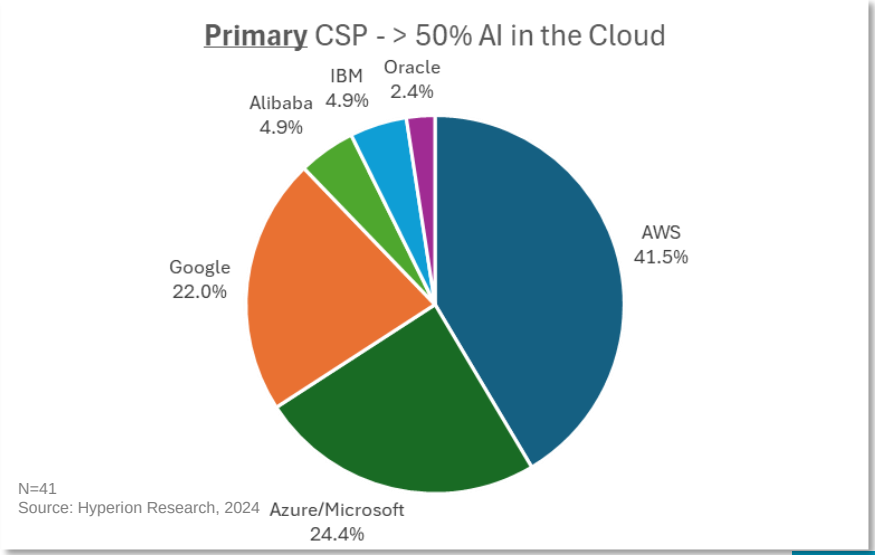


CSP Preferences – AI Workload Crosscut

Who is your PRIMARY cloud provider / ALL cloud providers for your HPC/AI/HPDA workloads TODAY?

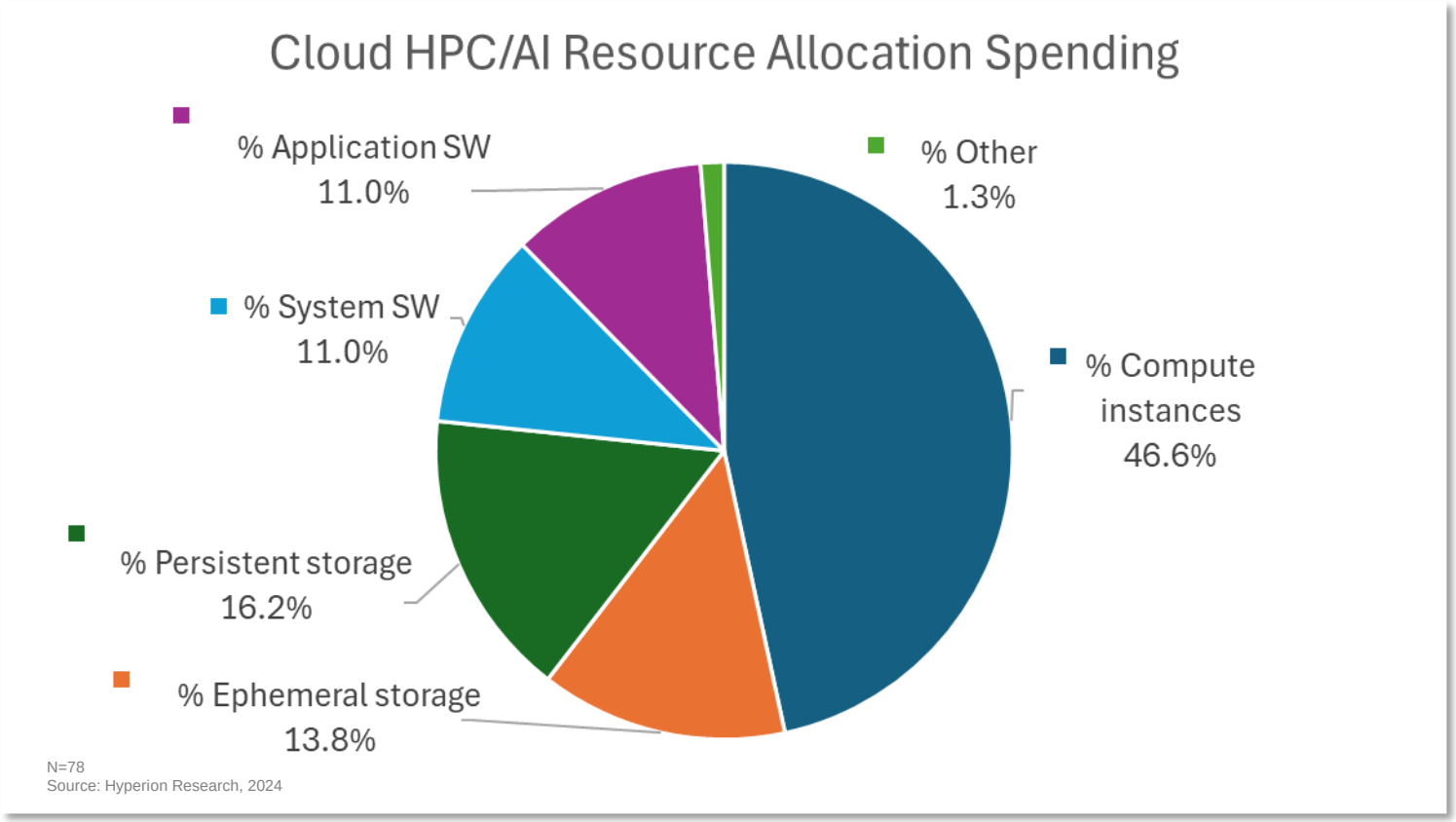
- **AWS the preferred primary CSP among respondents**
- **AWS as the primary CSP preference increases for sites who run >50% of their AI workloads in the cloud**
- **Microsoft moves to 2nd preferred primary preference for sites who run >50% of their AI workloads in the cloud**





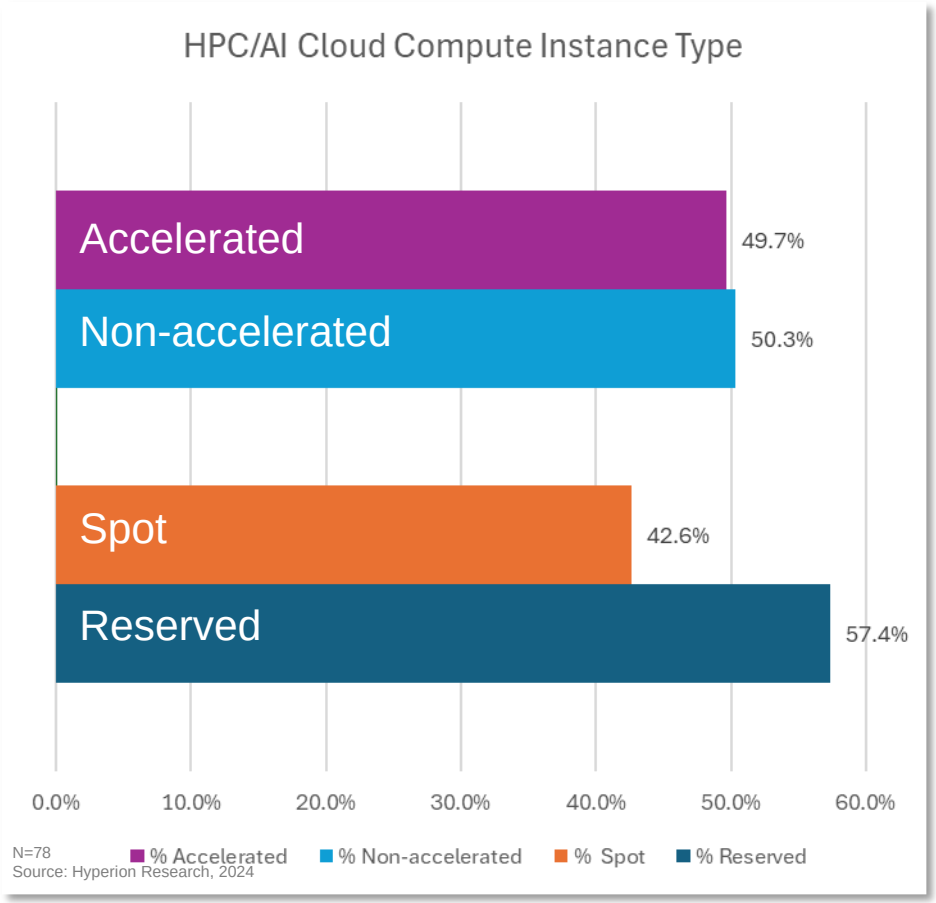
HPC-AI Cloud Resource Allocation Spending

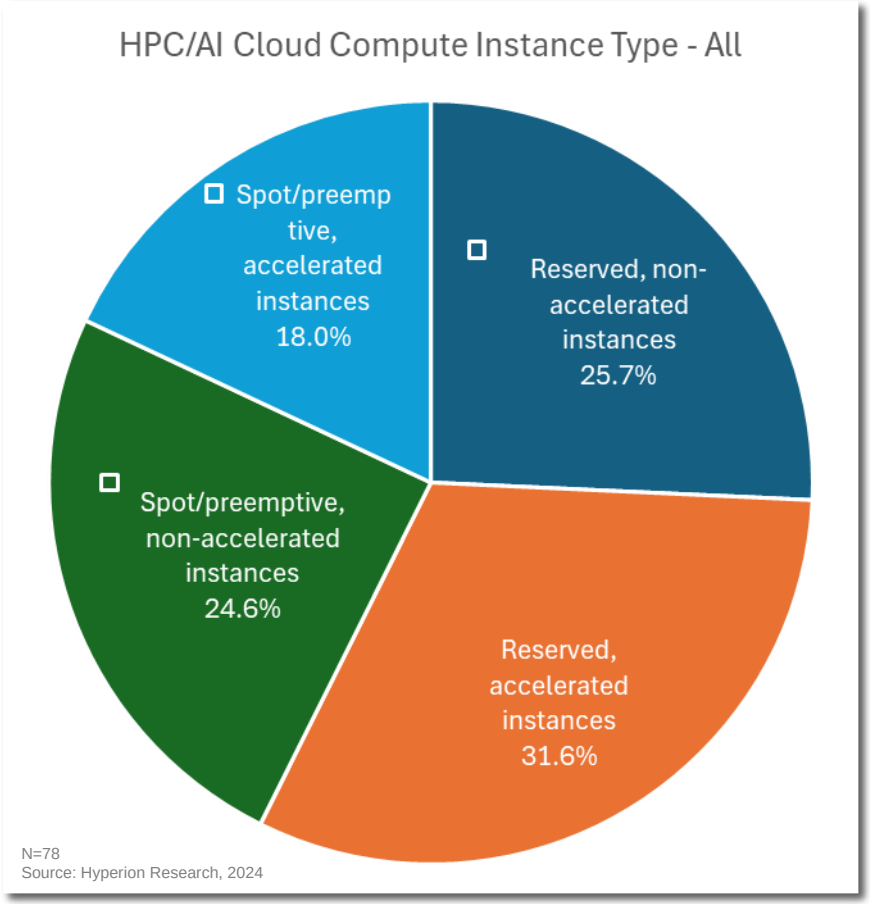
Please distribute your total HPC/AI/HPDA cloud resource spending between the following categories



HPC-AI Cloud Compute Instance Type

Please estimate your cloud instances for HPC/AI/HPDA workloads among the following categories:





What's Next in Clouds?

Dynamic environment anticipated for continuum computing and cloud adoption for the foreseeable future

- **Continued growth of user spending for HPC/AI advanced computing resources in the cloud from both new users and migration of workloads from current users**
- **More "specialization"**
  - Focused CSP accelerator and system designs for optimized performance and energy utilization
  - Capabilities (e.g., AI/GPU clouds) and outcomes (e.g., scope and complexity of workloads)
- **Build-out of new co-located hyperscaler data centers to support:**
  - New liquid cooling requirements of advanced architectures
  - Increasing number of "mid tier" service providers

HYPERION RESEARCH

Questions? We look forward to hearing from you!

mnossokoff@hyperionres.com